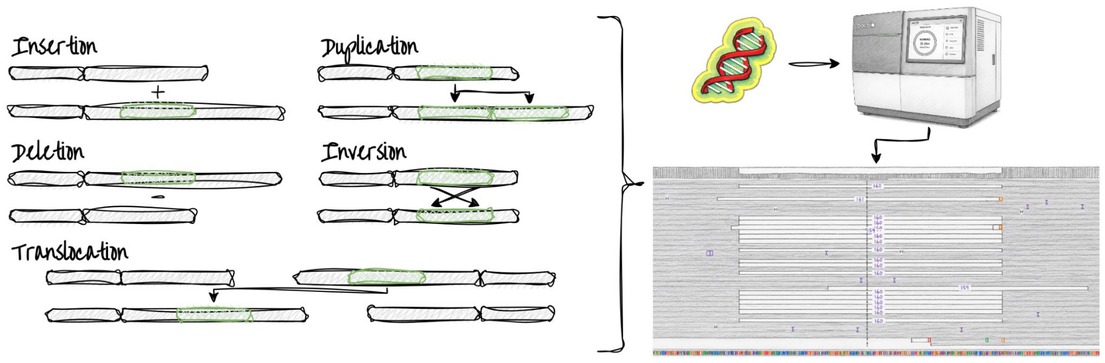
Somatic Structural Variant Discovery from Tumor Specific Strings

Introduction

Somatic structural variants (SVs) are key drivers of **cancer** but remain challenging to detect due to repetitive genomic regions, complex rearrangements, sequencing artifacts, and limited support for low-frequency events.

Existing approaches compare tumor and matched-normal reads against a **generic reference genome**, making it difficult to distinguish somatic variants from germline variation. These methods typically identify somatic variants through comparison of independently generated tumor and normal callsets, potentially **propagating errors** from both callsets.

Patient-specific germline assemblies enable direct tumor-germline comparison and improve somatic SV detection, particularly in complex genomic regions. However, these approaches depend on **high-quality, nearly complete diploid assemblies**, which currently require sequencing efforts beyond those of typical tumor-normal studies.

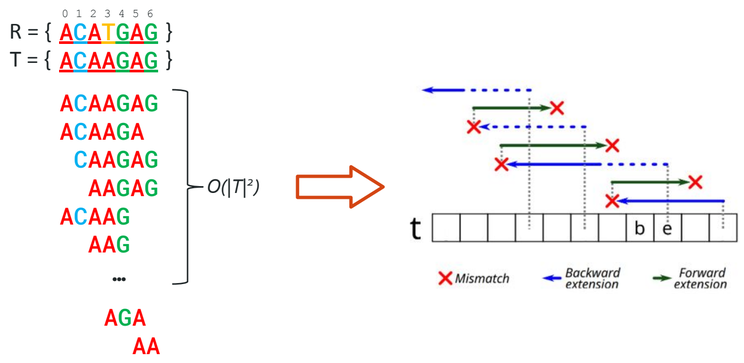


Terminology: *germline = inherited/normal; somatic = tumor-specific; calling = discovery/detection*

Specific strings

Comparative genomic analyses *(Khorsand et al. Bioinformatics Advances, 2021)*

Specific strings are the shortest strings in a target sample absent from a reference set, that can be efficiently computed using a FMD-Index. When applied to HiFi samples and trios, they capture >98% of the variation, including de novo SVs.



Germline SVs in hard-to-call regions *(Denti et al. Nature Methods, 2023)*

SVDSS outperforms mapping-based callers for insertions and deletions, with the largest gains in repetitive regions.

Pangenome graph augmentation *(Denti et al. bioRxiv, 2025)*

PALSS is the first assembly-free and mapping-free method for augmenting pangenome graphs directly from unassembled long reads, accurately recovering novel sequences without de novo assembly or variant calling.

Contribution

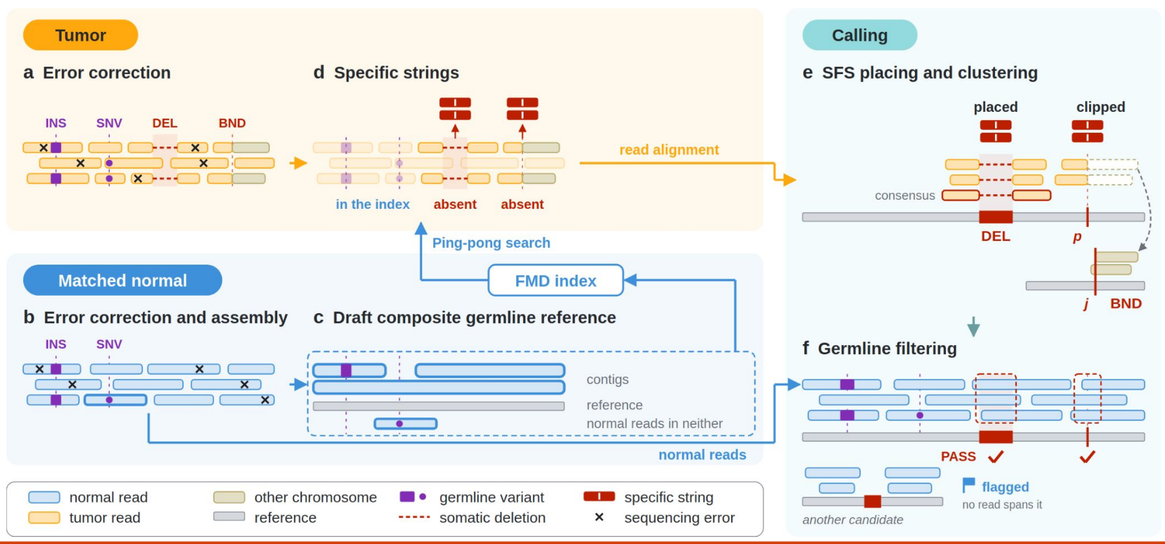
A specific string is the shortest tumor DNA sequence absent from the patient's germline genome, making it a **direct signal of tumor-specific variation**. Rather than relying on expensive, high-quality diploid assemblies, specific strings are computed against a **Draft Composite Germline Reference (DCGR)**, a reasonably complete representation of the patient's germline genome.

Rationale:

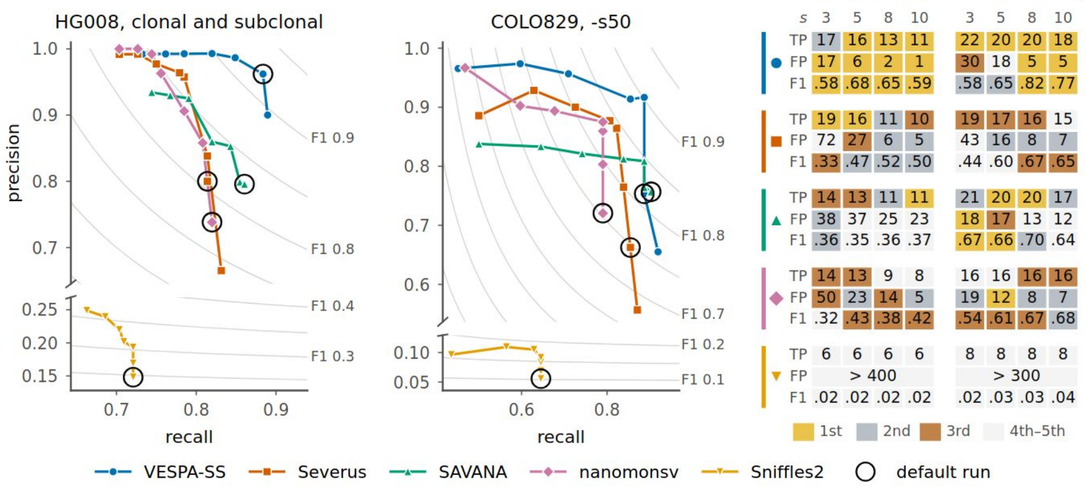
Germline variation represented in the DCGR does not generate specific strings, reducing the number of germline variants that reach the variant-calling stage

Key contributions:

- ✓ DCGR as a practical alternative to complete diploid germline assemblies
- ✓ Direct extraction of somatic SV evidence from tumor reads
- ✓ VESPA-SS, with state-of-the-art SV calling performance on benchmark datasets



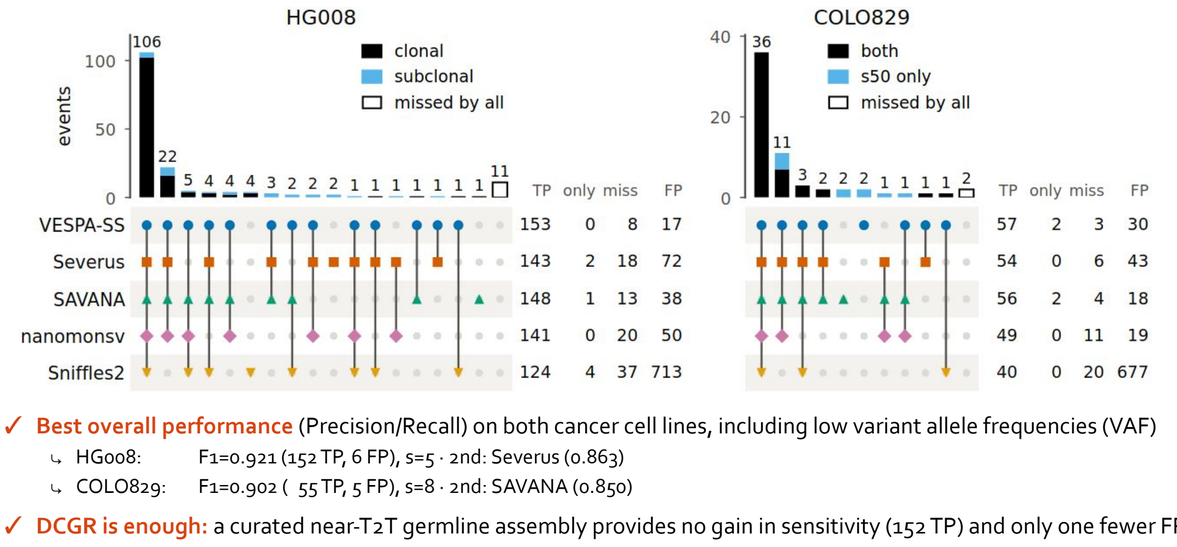
Results



events with VAF < 0.20

		HG008 (n=25)				COLO829 (n=24)			
	s	3	5	8	10	3	5	8	10
●	TP	17	16	13	11	22	20	20	18
	FP	17	6	2	1	30	18	5	5
	F1	.58	.68	.65	.59	.58	.65	.82	.77
■	TP	19	16	11	10	19	17	16	15
	FP	72	27	6	5	43	16	8	7
	F1	.33	.47	.52	.50	.44	.60	.67	.65
▲	TP	14	13	11	11	21	20	20	17
	FP	38	37	25	23	18	17	13	12
	F1	.36	.35	.36	.37	.67	.66	.70	.64
◆	TP	14	13	9	8	16	16	16	16
	FP	50	23	14	5	19	12	8	7
	F1	.32	.43	.38	.42	.54	.61	.67	.68
▼	TP	6	6	6	6	8	8	8	8
	FP		> 400				> 300		
	F1	.02	.02	.02	.02	.02	.03	.03	.04

1st 2nd 3rd 4th–5th



	TP	only	miss	FP
VESPA-SS	153	0	8	17
Severus	143	2	18	72
SAVANA	148	1	13	38
nanomonsv	141	0	20	50
Sniffles2	124	4	37	713

	TP	only	miss	FP
VESPA-SS	57	2	3	30
Severus	54	0	6	43
SAVANA	56	2	4	18
nanomonsv	49	0	11	19
Sniffles2	40	0	20	677

- ✓ **Best overall performance** (Precision/Recall) on both cancer cell lines, including low variant allele frequencies (VAF)
  - ↳ HG008: F1=0.921 (152 TP, 6 FP), s=5 · 2nd: Severus (0.863)
  - ↳ COLO829: F1=0.902 (55 TP, 5 FP), s=8 · 2nd: SAVANA (0.850)
- ✓ **DCGR is enough:** a curated near-T2T germline assembly provides no gain in sensitivity (152 TP) and only one fewer FP

Conclusions

Separating somatic from germline sequence before calling, *via specific strings against an inexpensive composite germline reference,* **improves somatic SV discovery, particularly for low-VAF events**

Ongoing work:

- ✓ Improving detection of complex rearrangements with shared breakpoints
- ✓ Extending evaluation to lower VAFs and lower tumor purities

https://github.com/dcmonti/VESPA

This work is supported by the European Union's Horizon programme under grant agreement No. 101180581 (ASVA-CGR)

FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY
Univerzita Komenského v Bratislave

MATFYZ CONNECTIONS

Luca Denti
Comenius University in Bratislava. Slovakia

Fereydoun Hormozdiari
University of California, Davis. USA

Paola Bonizzoni, Simone Ciccolella, Davide Cozzi, Gianluca Della Vedova, Davide Cesare Monti
University of Milano-Bicocca. Italy